CSE574 Introduction to Machine Learning

Discriminant Analysis

Jue Guo

University at Buffalo

March 26, 2024

Outline

Learning Objective

- Define discriminant function and linear discriminant analysis.
- Use discriminant functions to classify an instance.
- Define covariance.
- Use covariance and class means to calculate linear discriminant weights.
- Define principal components.
- Explain when the principal components technique is useful.

Linear Discriminant Analysis

A **discriminant function**, denoted $\delta(\mathbf{x})$, is a function used to set a decision boundary between classes.

- Each class has a unique discriminant function, $\delta_i(\mathbf{x})$, and the class with the highest discriminant function for a given set of input values is the predicted class.
- Ex: In logistic regression, the probabilities of class 1 vs. class 0 are discriminant functions $-\delta_1(\mathbf{x}) = \hat{p}_i = \frac{\exp(w_0+w_1x_i)}{1+\exp(w_0+w_1x_i)}$ vs. $\delta_0(\mathbf{x}) = 1 - \frac{\exp(w_0+w_1x_i)}{1+\exp(w_0+w_1x_i)}$.

If $\delta_1(\mathbf{x}) \geq \delta_0(\mathbf{x})$, then the instance is predicted as class 1 .

Linear Discriminant Analysis

Linear discriminant analysis classifies instances by comparing linear discriminant functions.

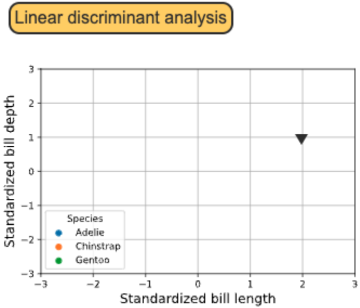
- Each discriminant function in linear discriminant analysis estimates the natural log of the conditional probability of class $i_1 P(y_i \mid \mathbf{x})$, based on a linear function of the input features.

$$\delta_i(\mathbf{x}) = \ln\left(P\left(y_i \mid \mathbf{x}\right)\right) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where $\mathbf{w}$ is a vector of weights for each input feature on class i, and w_{i0} is an intercept term for class i. The class with the highest discriminant function, $\delta_i(\mathbf{x}) = \ln\left(P\left(y_i \mid \mathbf{x}\right)\right)$, is the predicted class.

- The linear discriminant weights $\mathbf{w}$ are chosen so that the discriminant functions capture as much variation in the input features as possible while accurately classifying instances.

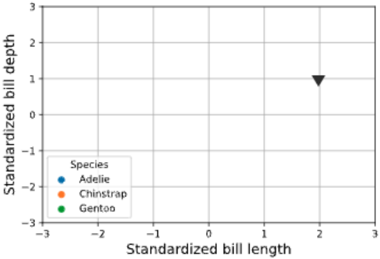
Classification using linear discriminant functions



Consider a penguin with standardized bill length 2 and standardized bill depth 1. Since three species are represented in the data, three discriminant functions will be used to classify the new penguin.

Classification using linear discriminant functions

Linear discriminant analysis



$$\delta_A(x_1, x_2) = -5.87x_1 + 4.80x_2 - 5.05$$

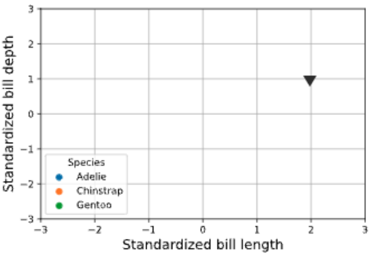
$$\delta_C(x_1, x_2) = 2.65x_1 + 0.62x_2 - 2.96$$

$$\delta_G(x_1, x_2) = 5.69x_1 - 6.25x_2 - 6.34$$

Let x_1 = standardized bill length and x_2 = standardized bill depth. $\delta_A(x_1, x_2)$, $\delta_C(x_1, x_2)$, and $\delta_G(x_1, x_2)$ denote the discriminant functions for Adelie, Chinstrap, and Gentoo species respectively.

Classification using linear discriminant functions

Linear discriminant analysis



$$\delta_A(x_1, x_2) = -5.87x_1 + 4.80x_2 - 5.05$$
$$\delta_A(2, 1) = -5.87 * 2 + 4.80 * 1 - 5.05$$

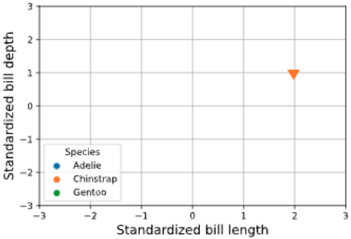
$$\delta_C(x_1, x_2) = 2.65x_1 + 0.62x_2 - 2.96$$
$$\delta_C(2, 1) = 2.65 * 2 + 0.62 * 1 - 2.96$$

$$\delta_G(x_1, x_2) = 5.69x_1 - 6.25x_2 - 6.34$$
$$\delta_G(2, 1) = 5.69 * 2 - 6.25 * 1 - 6.34$$

$x_1 = 2$ and $x_2 = 1$ is plugged into each discriminant function.

Classification using linear discriminant functions

Linear discriminant analysis



$$\delta_A(x_1, x_2) = -5.87x_1 + 4.80x_2 - 5.05$$
$$\delta_A(2, 1) = -5.87 * 2 + 4.80 * 1 - 5.05 = -11.99$$

$$\delta_C(x_1, x_2) = 2.65x_1 + 0.62x_2 - 2.96$$
$$\delta_C(2, 1) = 2.65 * 2 + 0.62 * 1 - 2.96 = \boxed{2.96}$$

$$\delta_G(x_1, x_2) = 5.69x_1 - 6.25x_2 - 6.34$$
$$\delta_G(2, 1) = 5.69 * 2 - 6.25 * 1 - 6.34 = -1.21$$

$\delta_C(2, 1)$ is the highest value, so the predicted class is Chinstrap.

Calculating Discriminant Weights using Covariance

Let $\mathbf{w}_i$ be the discriminant function weights for class i. The discriminant weights for standardized input features are calculated using two quantities:

- Class means for feature i, denoted $\mu_\mathbf{i}$.
- Covariance matrix, denoted $\mathbf{\Sigma}$.

Calculating Discriminant Weights using Covariance

- **Covariance** measures how values of one feature change in relation to a second feature, denoted σ_{ij}^2. Features with $\sigma_{ij}^2 = 0$ are uncorrelated, or independent. The variance between feature i and itself is the variance σ_i^2.
- A **covariance matrix** is a matrix containing all pairwise covariances between features i and j, denoted Σ.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \dots & \sigma_{1p}^2 \\ \sigma_{12}^2 & \sigma_2^2 & \dots & \sigma_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p}^2 & \sigma_{2p}^2 & \dots & \sigma_p^2 \end{bmatrix}$$

Calculating Discriminant Weights using Covariance

The discriminant function weights in linear discriminant analysis are

$$\mathbf{w}_i = \mathbf{\Sigma}^{-1}\mu_i$$

$$w_{0i} = -\frac{1}{2}\mu_i^T\mathbf{w}_i + \ln(P(y=i))$$

where $\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix and $P(y=i)$ is the prior probability of class i.

- The discriminant weights w_i are a weighted function of the class means.
- The intercept w_{0i} is a weighted function of the discriminant weights plus the natural log of the prior probability for class i, $P(y=i)$.
- The prior probability $P(y=i)$ may be estimated from the data or based on external assumptions.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

The class means μ_i represent the mean of each feature within class i. Ex: The mean standardized bill length for Adelie penguins is -0.9466.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

Covariance matrix, $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.2934 & 0.1633 \\ 0.1633 & 0.3236 \end{bmatrix}$$

The covariance matrix $\mathbf{\Sigma}$ contains the covariances for all pairs of features. Ex: The covariance between standardized bill length and standardized bill depth is 0.1633.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

Covariance matrix, $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.2934 & 0.1633 \\ 0.1633 & 0.3236 \end{bmatrix}$$

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 4.7393 & -2.3919 \\ -2.3919 & 4.2970 \end{bmatrix}$$

$$\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The inverse covariance matrix is the matrix such that $\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = I$, a matrix with 1 in the diagonal elements and 0 in the off-diagonal elements.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

Covariance matrix, $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.2934 & 0.1633 \\ 0.1633 & 0.3236 \end{bmatrix}$$

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 4.7393 & -2.3919 \\ -2.3919 & 4.2970 \end{bmatrix}$$

$$\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\boxed{w_1:}$ $w_1 = \mu_1 * \sigma_{11}^2 + \mu_2 * \sigma_{12}^2$

$$w_1 = -0.9466 * 4.7393 + 0.6013 * -2.3919$$

$$w_1 = -5.9244$$

For Adelie penguins, w_1 equals $\mu_1 * \sigma_{11}^2 + \mu_2 * \sigma_{12}^2 = -5.9244$.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

Covariance matrix, $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.2934 & 0.1633 \\ 0.1633 & 0.3236 \end{bmatrix}$$

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 4.7393 & -2.3919 \\ -2.3919 & 4.2970 \end{bmatrix}$$

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\boxed{w_1:}$ $w_1 = \mu_1 * \sigma_{11}^2 + \mu_2 * \sigma_{12}^2$

$$w_1 = -0.9466 * 4.7393 + 0.6013 * -2.3919$$

$$w_1 = -5.9244$$

$\boxed{w_2:}$ $w_2 = \mu_1 * \sigma_{12}^2 + \mu_2 * \sigma_{22}^2$

$$w_2 = -0.9466 * -2.3919 + 0.6013 * 4.2970$$

$$w_2 = 4.8480$$

w_2 equals $\mu_1 * \sigma_{12}^2 + \mu_2 * \sigma_{22}^2 = 4.840$.

Calculating discriminant weights for Adelie penguins.

Class means, μ_i

Class	μ_1	μ_2
Adelie	-0.9466	0.6013
Chinstrap	0.8866	0.6386
Gentoo	0.6548	-1.1027

Covariance matrix, $\mathbf{\Sigma}$

$$\mathbf{\Sigma} = \begin{bmatrix} 0.2934 & 0.1633 \\ 0.1633 & 0.3236 \end{bmatrix}$$

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 4.7393 & -2.3919 \\ -2.3919 & 4.2970 \end{bmatrix}$$

$$\mathbf{\Sigma}\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\boxed{w_1:}$ $w_1 = \mu_1 * \sigma^2_{11} + \mu_2 * \sigma^2_{12}$

$$w_1 = -0.9466 * 4.7393 + 0.6013 * -2.3919$$

$$w_1 = -5.9244$$

$\boxed{w_2:}$ $w_2 = \mu_1 * \sigma^2_{12} + \mu_2 * \sigma^2_{22}$

$$w_2 = -0.9466 * -2.3919 + 0.6013 * 4.2970$$

$$w_2 = 4.8480$$

$\boxed{w_0:}$ $w_0 = -\frac{1}{2}(\mu_1 * w_1 + \mu_2 * w_2) + \ln(P(Adelie))$

$$w_0 = -\frac{1}{2}(-0.9466 * -5.9244 + 0.6013 * 4.8480) + \ln\left(\frac{146}{333}\right)$$

$$w_0 = -5.0861$$

The intercept is $w_0 = -\frac{1}{2}(\mu_1 * w_1 + \mu_2 * w_2) + \ln(P(Adelie))$. Using the sample proportion 146/333 gives $w_0 = -5.0861$.

Practice Problem: Calculating discriminant weights.

Let x_1 = standardized flipper length and x_2 = standardized body mass. Use the following table of class means and inverse covariance matrix to calculate the discriminant weights for Gentoo penguins.

Class	μ_1	μ_2	n
Adelie	-0.7763	-0.6230	146
Chinstrap	-0.3765	-0.5895	68
Gentoo	1.1625	1.1012	119

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 6.7846 & -3.3188 \\ -3.3188 & 4.6956 \end{bmatrix}$$

1) $w_1 = \mu_1 * \sigma_{11}^2 + \mu_2 * \sigma_{12}^2 =$ ____
   - ○ $1.1625 * 6.7846 + 1.1012 * -3.3188$
   - ○ $1.1625 * -3.3188 + 1.1012 * 6.7846$
   - ○ $1.1012 * 6.7846 + 1.1625 * -3.3188$

2) w_2 = ____
   - ○ -1.5483
   - ○ -0.3489
   - ○ 1.3127

3) w_0 = ____
   - ○ -4.2119
   - ○ -2.1750
   - ○ -1.0290

Practice Problem: Calculating discriminant weights.

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 6.7846 & -3.3188 \\ -3.3188 & 4.6956 \end{bmatrix}$$

1) $w_1 = \mu_1 * \sigma_{11}^2 + \mu_2 * \sigma_{12}^2 =$ ____

- (selected) $1.1625 * 6.7846 + 1.1012 * -3.3188$
- $1.1625 * -3.3188 + 1.1012 * 6.7846$
- $1.1012 * 6.7846 + 1.1625 * -3.3188$

Correct

In the covariance matrix, $\sigma_{11}^2 = 6.7846$ and $\sigma_{12}^2 = -3.3188$. In the table of class means, $\mu_1 = 1.1625$ and $\mu_2 = 1.1012$. So, $w_1 = 1.1625 * 6.7846 + 1.1012 * -3.3188 = 4.2324$

2) w_2 = ____

- -1.5483
- -0.3489
- (selected) 1.3127

Correct

$w_2 = \mu_1 * \sigma_{12}^2 + \mu_2 * \sigma_{22}^2$. So, $w_2 = 1.1625 * -3.3188 + 1.1012 * 4.6956 = 1.3127$

3) w_0 = ____

- (selected) -4.2119
- -2.1750
- -1.0290

Correct

$w_0 = -\frac{1}{2}(\mu_1 * w_1 + \mu_2 * w_2) + \ln(P(Gentoo))$

So,

$w_0 = -\frac{1}{2}(1.1625 * 4.2324 + 1.1012 * 1.3127) + \ln(119/333) = -4.2119$

Principal Components

Linear discriminant analysis assumes that input features are uncorrelated, which is unrealistic in most cases.

- **Principal components** is a technique for creating linear combinations of input features that are uncorrelated, called the principal components. In principal components, a set of p input features is replaced by up to p linear combinations.

$$pc_j = \lambda_{1j}x_1 + \lambda_{2j}x_2 + \ldots + \lambda_{pj}x_p$$

Principal Components

In principal components, a set of p input features is replaced by up to p linear combinations.

$$pc_j = \lambda_{1j}x_1 + \lambda_{2j}x_2 + \ldots + \lambda_{pj}x_p$$

- The principal component weights, λ_{ij}, are chosen such that the resulting principal components pc_i and pc_j are uncorrelated, or independent.

Since pc_i and pc_j are independent, using principal components satisfies the linear discriminant assumption of uncorrelated input features. As a result, linear discriminant functions are often estimated based on the principal components rather than the original inputs.

First thing first, PCA only works for continous variables. With that in mind, let's continue with the examples.

Given a dataset with 2 features, X_1 and X_2, as follows:

Feature X_1	**Feature X_2**
2	1
4	3
6	5
8	7

How do we do PCA?

First, **standardize** each feature to have a mean of 0 and a standard deviation of 1.

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

Second, **compute the covariance matrix** Σ of the standardized features Z_1 and Z_2 :

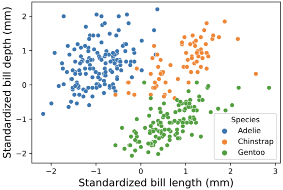
$$\Sigma = \frac{1}{n-1} \sum (Z - \bar{Z})(Z - \bar{Z})^T$$

Third, do a **eigendecomposition**, find eigenvalues (λ) and eigenvectors (v) of Σ, and the principal components, are represented by the eigenvectors of covariance matrix:

$$\Sigma v = \lambda v$$

Applying Principal Components.

Original input features

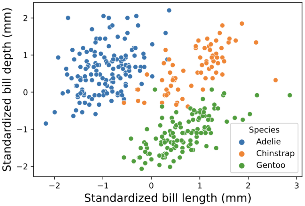


Principal components

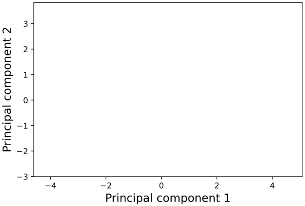
In the original feature space, standardized bill length (x_1) and standardized bill depth (x_2) have a positive correlation. As bill length increases, bill depth also increases.

Applying Principal Components.

Original input features



Principal components



Principal components

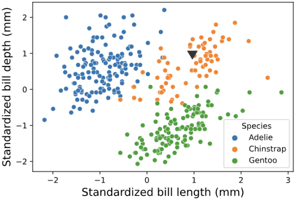
$$pc_1 = 1.942x_1 - 0.998x_2$$

$$pc_2 = -1.777x_1 - 1.049x_2$$

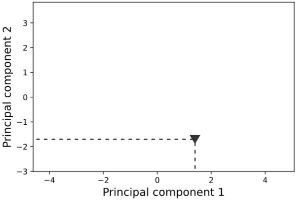
Two principal components are calculated to replace the original input features, pc_1 and pc_2.

Applying Principal Components.

Original input features



Principal components



Principal components

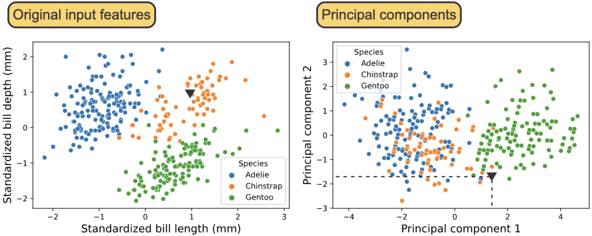
$x_1 = 1, x_2 = 1$

$pc_1 = 1.942x_1 - 0.998x_2 \longrightarrow pc_1 = 1.942 * 1 - 0.998 * 1 = 0.944$

$pc_2 = -1.777x_1 - 1.049x_2 \longrightarrow pc_2 = -1.777 * 1 - 1.049 * 1 = -2.826$

Each instance is transformed using the principal components. Ex: An instance at $(x_1 = 1, x_2 = 1)$ maps to $(pc_1 = 0.944, pc_2 = -2.826)$.

Applying Principal Components.



Principal components

$$x_1 = 1, x_2 = 1$$

$$pc_1 = 1.942x_1 - 0.998x_2 \longrightarrow pc_1 = 1.942 * 1 - 0.998 * 1 = 0.944$$

$$pc_2 = -1.777x_1 - 1.049x_2 \longrightarrow pc_2 = -1.777 * 1 - 1.049 * 1 = -2.826$$

Principal components pc_1 and pc_2 are independent, since no trend exists in the scatter plot.

Advantages and Disadvantages

Linear discriminant analysis extends to any number of classes, and the prior assumptions about class probabilities can be adjusted to better reflect reality or prior information.

- Ex: A previous study may have measured penguin populations at a different location, which could be incorporated into the prior probabilities. Using principal components as an intermediate step avoids issues with correlated input features. But, restricting the discriminant functions to linear equations results in a linear decision boundary.

Quadratic discriminant analysis uses quadratic equations in the discriminant functions. The resulting discriminant equations are more complicated, but in some situations a curved decision boundary is a better fit. A tradeoff exists between model complexity and interpretability—models that are more complex are less interpretable.